Contributions to Mokken's nonparametric item response theory

Sijtsma, Klaas

Summary

In this study, the item response models proposed by Mokken (1971) are discussed, and further developed. In these item response models the relation between observable response behavior on an item and latent characteristics of persons and items is described by means of an item characteristic function. Since these functions are not defined parametrically in Mokken's approach, the item response models are called nonparametric. The models proposed by Mokken can be used to scale persons and items on a single dimension. This dimension represents the quantitative measurement scale of a psychological or some other attribute. By measuring persons by means of tests or questionnaires which comply with an item response model, assertions about behavior are psychometrically well founded.

In Chapter One, an overview is given of test models that are currently used for test construction. Mokken's models and results obtained in the field of nonparametric Item Response Theory are discussed extensively. Two models are of special interest in this study. The first is the model of monotone homogeneity, that leads to an order of persons along the latent measurement scale. The second is the model of double monotonicity, which allows an order of both persons and items. The latter model is a special case of the former. Theoretical and observable properties of both models as well as some of their empirical applications to test data are discussed.

Four different subjects pertaining to Mokken's models are treated in chapters Two through Five. Each chapter is devoted to one subject. Extensions, refinements and additional results are presented which further improve the usefulness of Mokken's approach in test construction research.

The goodness of fit of the models of monotone homogeneity and double monotonicity to empirical data is the subject of the second chapter. Mokken uses Loevinger's coefficient of scalability, H, and related item coefficients to assess the degree to which the model of monotone homogeneity fits the data. He further presents a visual inspection method to assess the fit of the model of double monotonicity. In Chapter Two, an alternative method is presented to study goodness of fit. By means of this alternative method it is possible to assess the goodness of fit of both nonparametric models to empirical data. Furthermore, individual

items can be assessed with respect to their fit, thus making possible decisions about acceptance or rejection of items. By means of a Monte Carlo study, the behavior of this alternative method across samples is compared with the behavior of the H coefficient. Their capability to find items violating an item response model is specifically considered. The sampling behavior of the newly proposed method is studied, and rules of thumb are provided to facilitate its practical use. Both methods prove to be useful tools to assess the quality of individual, as well as sets of, items. It is also clear that both methods are susceptible to partly different properties of a data set, and, therefore, they can not be used interchangeably.

If items have been selected by means of the methods treated in Chapter Two, then the resulting set does not necessarily have high discriminating power. High discriminating power is better attained if the data tend to comply with the Guttman model, indicated by a high H value. The subject of Chapter Three is the stepwise selection of items by means of a bottom up procedure. This procedure seeks to achieve a high H value for the complete set. Starting with the pair of items having the highest H value, an item which maximizes the overall H coefficient of the subset of items already selected is added in each subsequent step. The properties of this selection procedure are studied in this chapter. Specifically, the order in which items are selected from an existing pool relative to their difficulty is investigated. Furthermore, the behavior of H is studied in the course of the bottom up procedure. Chapter Three also contains a systematic comparison of the H coefficient with the lower bound to the reliability known as coefficient alpha. From this comparison it is clear that these coefficients can not be used as indicators of the same properties of an item set.

Once a set of items has been selected to be the final test, the practical usefulness of the test as a measurement device is not guaranteed. In order to compare the measurement values of different persons with each other, or with an external criterion, these values should be accurate estimates of the person parameters. In other words, measurement should be reliable. In Chapter Four, two methods proposed by Mokken to estimate the reliability are studied, and a third is developed in the present study. These three methods assume that the model of double monotonicity holds.

Several refinements of the methods are proposed. By means of a Monte Carlo study, the sampling behavior of these three methods is compared with the sampling behavior of four methods from Classical Test Theory. An important result of this study is that one of Mokken's methods, and the method developed in Chapter Four, lead to almost unbiased estimates of the reliability. Their sampling variance is approximately equal to the variance of the classical estimates.

A simple measurement value for each person may be regarded insufficient in some applications. Sometimes, the pattern of item scores may reveal additional information about an examinee useful for diagnostic purposes. In Chapter Five, a scalability coefficient for persons is proposed. This coefficient can be viewed as the person counterpart of the item coefficient of Mokken. The person coefficient expresses the degree to which a given item response pattern is comparable with the patterns of other persons. In this way, deviant patterns can be found. Several characteristics of this coefficient are derived. From a simulation study, it can be concluded that the coefficient is rather successful in detecting a minority of aberrant patterns among a majority of patterns complying with the same underlying item response model.